Linear Discriminant Analysis

Nate Wells

Math 243: Stat Learning

November 3rd, 2021

Outline

In today's class, we will. . .

- Discuss LDA theory and motivation
- Build an LDA classifier by hand

Section 1

LDA

Logistic Regression, KNN, and Bayes' Classifier

Recall that for a binary classification problem, the average test error rate is minimized using the Bayes' classifier:

$$f(x_0) = \text{argmax}_j P(Y = j \mid X = x_0) \quad j \in \{0, 1\}$$

Logistic Regression, KNN, and Bayes' Classifier

Recall that for a binary classification problem, the average test error rate is minimized using the Bayes' classifier:

$$f(x_0) = \mathrm{argmax}_j P(Y = j \mid X = x_0) \quad j \in \{0, 1\}$$

Both KNN and Logistic regression attempt to estimate the conditional probability $p(X) = P(Y = 1 \mid X)$:

Logistic Regression, KNN, and Bayes' Classifier

Recall that for a binary classification problem, the average test error rate is minimized using the Bayes' classifier:

$$f(x_0) = \mathrm{argmax}_j P(Y = j \mid X = x_0) \quad j \in \{0, 1\}$$

Both KNN and Logistic regression attempt to estimate the conditional probability $p(X) = P(Y = 1 \mid X)$:

- Logistic regression:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Logistic Regression, KNN, and Bayes' Classifier

Recall that for a binary classification problem, the average test error rate is minimized using the Bayes' classifier:

$$f(x_0) = \mathrm{argmax}_j P(Y = j \mid X = x_0) \quad j \in \{0, 1\}$$

Both KNN and Logistic regression attempt to estimate the conditional probability $p(X) = P(Y = 1 \mid X)$:

- Logistic regression:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

- KNN:

$$p(X) = \frac{1}{K} \sum_{i \in N_0} I(y_i = 1)$$

The Law of Total Probability

Suppose $A_1, A_2, \ldots, A_k$ are a list of events that are:

- *mutually exclusive*: $P(A_i \text{ and } A_j) = 0$
- *exhaustive*: $P(A_1) + P(A_2) \cdots + P(A_k) = 1$
  - Example: Flip two coins, and let $A_1 = \text{both flips are different}$, $A_2 = \text{both flips are heads}$, $A_3 = \text{both flips are tails}$.

Then for any other event B,

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)$$

The Law of Total Probability

Suppose $A_1, A_2, \dots, A_k$ are a list of events that are:

- *mutually exclusive*: $P(A_i \text{ and } A_j) = 0$
- *exhaustive*: $P(A_1) + P(A_2) \cdots + P(A_k) = 1$
    - Example: Flip two coins, and let $A_1 = \text{both flips are different}$, $A_2 = \text{both flips are heads}$, $A_3 = \text{both flips are tails}$.

Then for any other event B,

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)$$

Example

Consider two boxes of marbles, the first containing 60% blue and 40% red, and the second containing 10% blue and 90% red. Suppose we draw a marble from the first box with 20% probability and from the second box with 80% probability.

- What is the probability we draw a blue marble?

Bayes' Rule

For any events A and B,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Rule

For any events A and B,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A)$ is called the *prior probability* of A and represents our initial beliefs about the event A.
- Suppose B is an event that we observe occurring.
- $P(A|B)$ is called the *posterior probability* of A and represents our updated beliefs about the event A in light of the event B.

Bayes' Rule

For any events A and B,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A)$ is called the *prior probability* of A and represents our initial beliefs about the event A.
- Suppose B is an event that we observe occurring.
- $P(A|B)$ is called the *posterior probability* of A and represents our updated beliefs about the event A in light of the event B.

Example

Suppose a test for a certain disease has specificity .8 and sensitivity .95, and that the disease has prior prevalence of 0.01. Find the posterior probability that an individual who tests positive for the disease actually has the disease.

The Bayesian Flip

For classification problems, we want to know $P(Y = A_j \mid X = x_0)$.

The Bayesian Flip

For classification problems, we want to know $P(Y = A_j \,|\, X = x_0)$.

- Using Bayes' Rule:

$$P(Y = A_j \,|\, X = x_0) = \frac{P(X = x_0 \,|\, Y = A_j)P(Y = A_j)}{P(X = X_0)}$$
$$= \frac{P(X = x_0 \,|\, Y = A_j)P(Y = A_j)}{\sum_i P(X = X_0 | Y = A_i)P(Y = A_i)}$$

The Bayesian Flip

For classification problems, we want to know $P(Y = A_j \mid X = x_0)$.

- Using Bayes' Rule:

$$\begin{aligned} P(Y = A_j \mid X = x_0) &= \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{P(X = X_0)} \\ &= \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{\sum_i P(X = X_0 | Y = A_i)P(Y = A_i)} \end{aligned}$$

- We estimate the conditional probability of the response using...
  - The conditional distribution $P(X = x_0 \mid Y = A_j)$ of each predictor **given the response**
  - The prior distribution $\pi_i = P(Y = A_i)$ of the response

The Bayesian Flip

For classification problems, we want to know $P(Y = A_j \mid X = x_0)$.

- Using Bayes' Rule:

$$P(Y = A_j \mid X = x_0) = \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{P(X = X_0)}$$
$$= \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{\sum_i P(X = X_0 | Y = A_i)P(Y = A_i)}$$

- We estimate the conditional probability of the response using...
  - The conditional distribution $P(X = x_0 \mid Y = A_j)$ of each predictor **given the response**
  - The prior distribution $\pi_i = P(Y = A_i)$ of the response
- In practice, we don't have access to the conditional distributions of the predictors, so need to estimate them based on data.

LDA

- Suppose we have just one predictor X and a multi-level categorical response Y.

LDA

- Suppose we have just one predictor X and a multi-level categorical response Y.
- What is the most "natural" assumption for the conditional distribution of X, given $Y = A_j$?

LDA

- Suppose we have just one predictor X and a multi-level categorical response Y.
- What is the most "natural" assumption for the conditional distribution of X, given $Y = A_j$?
- If X is normal with mean μ_j and variance σ_j^2, its density is

$$P(X = x \mid Y = A_j) = f_j(x) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(x-\mu_j)^2/2\sigma_j^2}$$

LDA

- Suppose we have just one predictor X and a multi-level categorical response Y.
- What is the most "natural" assumption for the conditional distribution of X, given $Y = A_j$?
- If X is normal with mean μ_j and variance σ_j^2, its density is

$$P(X = x \mid Y = A_j) = f_j(x) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(x-\mu_j)^2/2\sigma_j^2}$$

- Moreover, if we assume all conditional distributions have the **same** variance $\sigma_j^2 = \sigma^2$, we can simplify our model.

Likelihood Ratio

- To determine to which class an observation belongs, based on the conditional distribution of predictors, we consider the likelihood ratio (LR):

$$\mathrm{LR} = \frac{P(Y = A_j | X = x_0)}{P(Y = A_k | X = x_0)}$$

Likelihood Ratio

- To determine to which class an observation belongs, based on the conditional distribution of predictors, we consider the likelihood ratio (LR):

$$\mathrm{LR} = \frac{P(Y = A_j | X = x_0)}{P(Y = A_k | X = x_0)}$$

- If $\mathrm{LLR} \geq 1$, we should predict A_j over A_k. Otherwise, predict A_k over A_j.

Likelihood Ratio

- To determine to which class an observation belongs, based on the conditional distribution of predictors, we consider the likelihood ratio (LR):

$$\mathrm{LR} = \frac{P(Y = A_j|X = x_0)}{P(Y = A_k|X = x_0)}$$

- If $\mathrm{LLR} \geq 1$, we should predict A_j over A_k. Otherwise, predict A_k over A_j.
- And using Bayes' Rule:

$$\begin{aligned}\frac{P(Y = A_j \mid X = x_0)}{P(Y = A_k \mid X = x_0)} &= \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)/P(X = x_0)}{P(X = x_0 \mid Y = A_k)P(Y = A_k)/P(X = x_0)} \\ &= \frac{P(X = x_0 \mid Y = A_j)P(Y = A_j)}{P(X = x_0 \mid Y = A_k)P(Y = A_k)} \\ &= \frac{e^{-(x_0-\mu_j)^2/2\sigma^2}\pi_j}{e^{-(x_0-\mu_k)^2/2\sigma^2}\pi_k}\end{aligned}$$

The Log-liklihood Ratio

The log-liklihood ratio is obtained by taking natural log of the liklihood ratio:

$$
\begin{aligned}
\ln \mathrm{LR} =& \ln \frac{P(Y = A_j \mid X = x_0)}{P(Y = A_k \mid X = x_0)} \\
=& \ln \frac{e^{-(x_0 - \mu_j)^2/2\sigma^2} \pi_j}{e^{-(x_0 - \mu_k)^2/2\sigma^2} \pi_k} \\
=& (x_0 - \mu_k)^2/2\sigma^2 - (x_0 - \mu_j)^2/2\sigma^2 + \ln \pi_j - \ln \pi_k
\end{aligned}
$$

The Log-liklihood Ratio

The log-liklihood ratio is obtained by taking natural log of the liklihood ratio:

$$\begin{aligned}\ln \mathrm{LR} &= \ln \frac{P(Y = A_j \mid X = x_0)}{P(Y = A_k \mid X = x_0)} \\ &= \ln \frac{e^{-(x_0-\mu_j)^2/2\sigma^2}\pi_j}{e^{-(x_0-\mu_k)^2/2\sigma^2}\pi_k} \\ &= (x_0-\mu_k)^2/2\sigma^2 - (x_0-\mu_j)^2/2\sigma^2 + \ln \pi_j - \ln \pi_k\end{aligned}$$

- The decision boundary between A_j and A_k is the point c where $\ln \mathrm{LR} = 0$, or

$$(c-\mu_k)^2/2\sigma^2 + \ln \pi_j = (c-\mu_j)^2/2\sigma^2 + \ln \pi_k$$

The Log-liklihood Ratio

The log-liklihood ratio is obtained by taking natural log of the liklihood ratio:

$$\begin{aligned}
\ln \mathrm{LR} &= \ln \frac{P(Y = A_j \mid X = x_0)}{P(Y = A_k \mid X = x_0)} \\
&= \ln \frac{e^{-(x_0-\mu_j)^2/2\sigma^2}\pi_j}{e^{-(x_0-\mu_k)^2/2\sigma^2}\pi_k} \\
&= (x_0 - \mu_k)^2/2\sigma^2 - (x_0 - \mu_j)^2/2\sigma^2 + \ln \pi_j - \ln \pi_k
\end{aligned}$$

- The decision boundary between A_j and A_k is the point c where $\ln \mathrm{LR} = 0$, or

$$(c - \mu_k)^2/2\sigma^2 + \ln \pi_j = (c - \mu_j)^2/2\sigma^2 + \ln \pi_k$$

- Solving for c gives

$$c = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2(\ln \pi_k - \ln \pi_j)}{\mu_j - \mu_k}$$

Binary Classfication with Uniform Prior

Suppose Y is binary, and that each of $X|Y=0$ and $X|Y=1$ are Normal with common variance σ and means μ_0 and μ_1. Moreover, assume a uniform prior $\pi_0 = \pi_1 = \frac{1}{2}$

Binary Classfication with Uniform Prior

Suppose Y is binary, and that each of $X|Y = 0$ and $X|Y = 1$ are Normal with common variance σ and means μ_0 and μ_1. Moreover, assume a uniform prior $\pi_0 = \pi_1 = \frac{1}{2}$

Solve for c in

$$(c - \mu_k)^2/2\sigma^2 + \ln \pi_j = (c - \mu_j)^2/2\sigma^2 + \ln \pi_k$$

Binary Classfication with Uniform Prior

Suppose Y is binary, and that each of $X|Y=0$ and $X|Y=1$ are Normal with common variance σ and means μ_0 and μ_1. Moreover, assume a uniform prior $\pi_0 = \pi_1 = \frac{1}{2}$

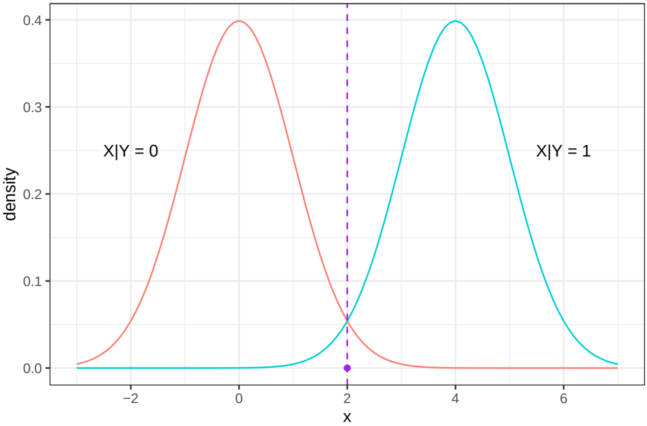
Solve for c in

$$(c-\mu_k)^2/2\sigma^2 + \ln \pi_j = (c-\mu_j)^2/2\sigma^2 + \ln \pi_k$$

We get $c = \frac{\mu_1+\mu_2}{2}$

Plots

Suppose $X|Y = 0 \sim N(0,1)$ and $X|Y = 1 \sim N(4,1)$



What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.
- A normal distribution requires only 2 parameters: μ and σ.

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.
- A normal distribution requires only 2 parameters: μ and σ.
  - We need one estimate of μ for each level of Y.
  - Since we assumed each conditional distribution had the same variance, we need only 1 estimate for σ

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.
- A normal distribution requires only 2 parameters: μ and σ.
  - We need one estimate of μ for each level of Y.
  - Since we assumed each conditional distribution had the same variance, we need only 1 estimate for σ
- LDA is an algorithm for obtaining these estimates and then classifying based on log-likelihood ratio.

What is LDA?

If we **knew** the conditional distribution of the predictors, we could easily create decision boundaries.

- But we only have data, so we need to estimate those distributions.
- A normal distribution requires only 2 parameters: μ and σ.
  - We need one estimate of μ for each level of Y.
  - Since we assumed each conditional distribution had the same variance, we need only 1 estimate for σ
- LDA is an algorithm for obtaining these estimates and then classifying based on log-likelihood ratio.
- Our estimates for μ_j and σ^2 are:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i:y_i=A_j} x_i \qquad \hat{\sigma}^2 = \frac{1}{n-\ell} \sum_{j=1}^{\ell} \sum_{i:y_i=A_j} (x_i - \hat{\mu}_j)^2$$

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

- The discriminant is obtained by taking log-probabilities and discarding terms in the sum that don't depend on j.

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

- The discriminant is obtained by taking log-probabilities and discarding terms in the sum that don't depend on j.
- We can then assign an observation x_0 to the class whose discriminant is largest at $x = x_0$.

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

- The discriminant is obtained by taking log-probabilities and discarding terms in the sum that don't depend on j.
- We can then assign an observation x_0 to the class whose discriminant is largest at $x = x_0$.
- Why is LDA called **Linear** Discriminant Analysis?

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

- The discriminant is obtained by taking log-probabilities and discarding terms in the sum that don't depend on j.
- We can then assign an observation x_0 to the class whose discriminant is largest at $x = x_0$.
- Why is LDA called **Linear** Discriminant Analysis?
  - Because the discriminant function is linear in x.

The Discriminant

Rather than comparing log likelihoods, we could instead look at the log conditional probability for each level. This function $\delta_j(x)$ is called the *discriminant* for level j:

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

- The discriminant is obtained by taking log-probabilities and discarding terms in the sum that don't depend on j.
- We can then assign an observation x_0 to the class whose discriminant is largest at $x = x_0$.
- Why is LDA called **Linear** Discriminant Analysis?
  - Because the discriminant function is linear in x.
  - Using this classification algorithm will result in linear decision boundaries.

Section 2

Handmade LDA model

LDA

Suppose Y is a categorical variable with ℓ levels, and for each level A_j, that

$$X|Y = A_j \sim N(\mu_j, \sigma).$$

LDA

Suppose Y is a categorical variable with ℓ levels, and for each level A_j, that

$$X|Y = A_j \sim N(\mu_j, \sigma).$$

The discriminant function

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

can be used to classify an observation by choosing the level A_j whose discriminant is largest at x.

LDA

Suppose Y is a categorical variable with ℓ levels, and for each level A_j, that

$$X|Y = A_j \sim N(\mu_j, \sigma).$$

The discriminant function

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

can be used to classify an observation by choosing the level A_j whose discriminant is largest at x.

We estimate the values of μ_j and σ from the sample data:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i:y_i = A_k} x_i$$

LDA

Suppose Y is a categorical variable with ℓ levels, and for each level A_j, that

$$X|Y = A_j \sim N(\mu_j, \sigma).$$

The discriminant function

$$\delta_j(x) = x \cdot \frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \ln \pi_j$$

can be used to classify an observation by choosing the level A_j whose discriminant is largest at x.

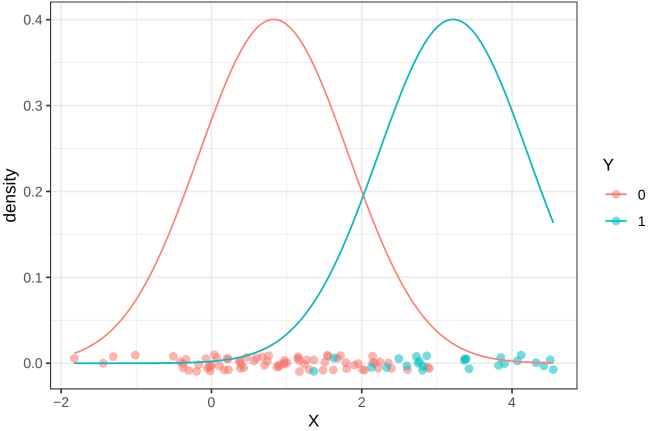
We estimate the values of μ_j and σ from the sample data:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i:y_i=A_k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-\ell} \sum_{j=1}^{\ell} \sum_{i:y_i=A_k} (x_i - \hat{\mu}_j)^2$$

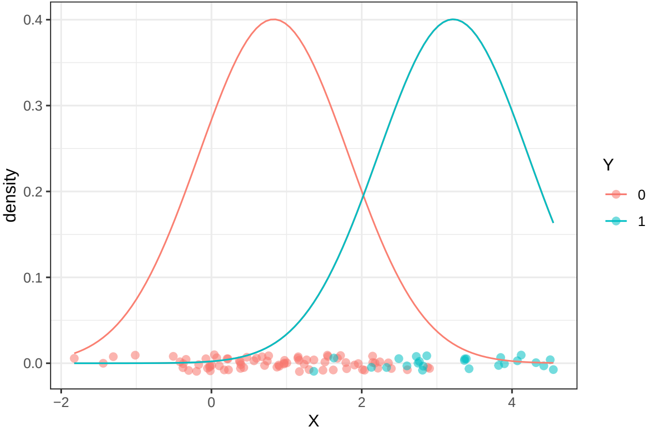
Simulated Data

Suppose $X|Y = 0 \sim N(1, 1)$ and $X|Y = 1 \sim N(3, 1)$, and that $\pi_0 = .75$ and $\pi_1 = .25$.



Simulated Data

Suppose $X|Y = 0 \sim N(1, 1)$ and $X|Y = 1 \sim N(3, 1)$, and that $\pi_0 = .75$ and $\pi_1 = .25$.



- What feature of the graph shows that $\pi_0 = .75$ and $\pi_1 = .25$?

Find Estimates

Estimates for μ_j and π_j

```
d %>% group_by(Y) %>% summarize(pi = n()/n, mu = mean(X))
```

```
## # A tibble: 2 x 3
##   Y        pi    mu
##   <fct> <dbl> <dbl>
## 1 0      0.75 0.828
## 2 1      0.25 3.22
```

Find Estimates

Estimates for μ_j and π_j

```
d %>% group_by(Y) %>% summarize(pi = n()/n, mu = mean(X))
```

```
## # A tibble: 2 x 3
##   Y        pi    mu
##   <fct> <dbl> <dbl>
## 1 0      0.75 0.828
## 2 1      0.25 3.22
```

Estimate for σ^2.

```
d %>% group_by(Y) %>% summarize(ssx = var(X) * (n() - 1)) %>%
  summarize(sigma_sq = sum(ssx)/(n-2))
```

```
## # A tibble: 1 x 1
##   sigma_sq
##      <dbl>
## 1    0.992
```

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

$$c = \frac{\mu_0 + \mu_1}{2} + \frac{\sigma^2(\ln \pi_0 - \ln \pi_1)}{\mu_1 - \mu_0}$$

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

$$c = \frac{\mu_0 + \mu_1}{2} + \frac{\sigma^2(\ln \pi_0 - \ln \pi_1)}{\mu_1 - \mu_0}$$

```
c<- (mu0 + mu1)/2 + (sigma2*log(pi0) - log(pi1))/(mu1-mu0)
c
```

```
## [1] 2.483001
```

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

$$c = \frac{\mu_0 + \mu_1}{2} + \frac{\sigma^2(\ln \pi_0 - \ln \pi_1)}{\mu_1 - \mu_0}$$

```
c<- (mu0 + mu1)/2 + (sigma2*log(pi0) - log(pi1))/(mu1-mu0)
c
```

```
## [1] 2.483001
```

Write a function to create discriminant functions:

```
discriminant <- function(x, pi, mu, sigma2) {
  x * (mu/sigma2) - (mu^2)/(2 * sigma2) + log(pi)
}
```

The discriminant function

Solve for intersection of discriminant functions: $\delta_0(c) = \delta_1(c)$ when

$$c = \frac{\mu_0 + \mu_1}{2} + \frac{\sigma^2(\ln \pi_0 - \ln \pi_1)}{\mu_1 - \mu_0}$$

```
c<- (mu0 + mu1)/2 + (sigma2*log(pi0) - log(pi1))/(mu1-mu0)
c
```

```
## [1] 2.483001
```

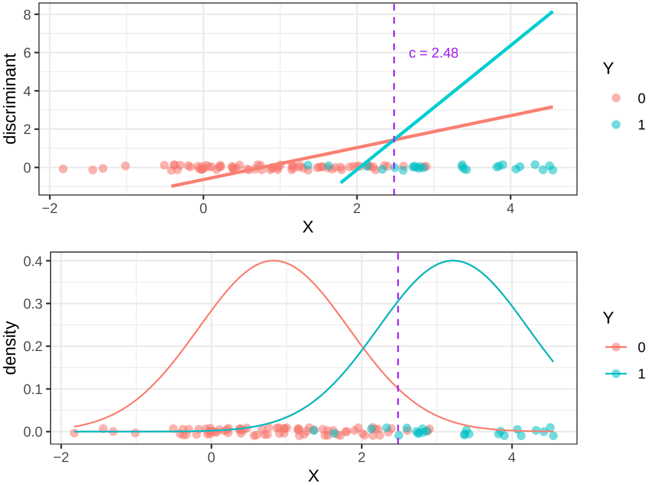
Write a function to create discriminant functions:

```
discriminant <- function(x, pi, mu, sigma2) {
  x * (mu/sigma2) - (mu^2)/(2 * sigma2) + log(pi)
}
```

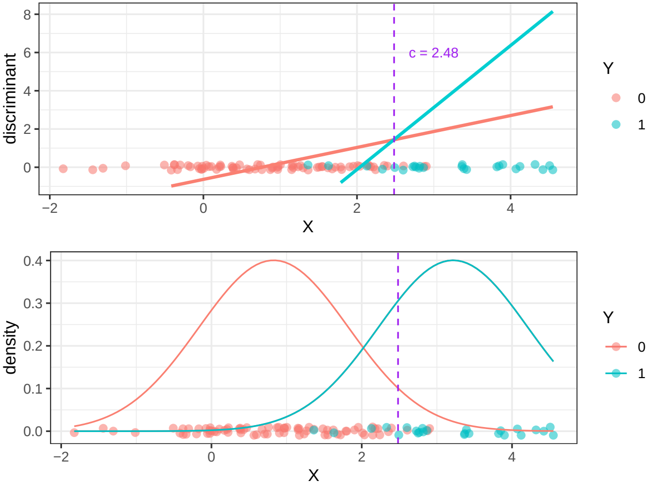
Evaluate discriminant function on data for each class:

```
d0 <- discriminant(d$X, pi0, mu0, sigma2)
d1 <- discriminant(d$X, pi1, mu1, sigma2)
```

Plots


8
6
4
2
0
discriminant
c = 2.48
−2
0
2
4
X
Y
0
1
0.4
0.3
0.2
0.1
0.0
density
−2
0
2
4
X
Y
0
1

Plots



- Why don't discriminant functions intersect at the same point as density curves?